AN ECONOMIST'S PRIMER ON SURVEY SAMPLES*

by

William J. Carrington
Welch Consulting and Unicon Research Corporation

John L. Eltinge
Bureau of Labor Statistics

and

Kristin McCue
Center for Economic Studies
U.S. Bureau of the Census

CES 00-15 October, 2000

ABSTRACT

Survey data underlie most empirical work in economics, yet economists typically have little familiarity with survey sample design and its effects on inference. This paper describes how sample designs depart from the simple random sampling model implicit in most econometrics textbooks, points out where the effects of this departure are likely to be greatest, and describes the relationship between design-based estimators developed by survey statisticians and related econometric methods for regression. Its intent is to provide empirical economists with enough background in survey methods to make informed use of design-based estimators. It emphasizes surveys of households (the source of most public-use files), but also considers how surveys of businesses differ. Examples from the National Longitudinal Survey of Youth of 1979 and the Current Population Survey illustrate practical aspects of design-based estimation.

Keywords: survey samples; variance estimation; clustering; weights.

JEL classification: C42, C81

* We are grateful for helpful comments received from seminar participants at the Center for Economic Studies. The views expressed are those of the authors and do not necessarily reflect the views or policies of the Census Bureau, the BLS, or any other agency of the U.S. Department of Labor.

1. Introduction

Survey data are used in most empirical work in economics, but economists typically have not considered survey sampling methods to be relevant to their analyses. This has begun to change as general purpose statistical packages have added estimators intended for use with complex samples, and as the variables needed to account for the effects of sample design have become more widely available. Yet economists' use of survey data continues to be hampered by a lack of familiarity with sample design and design-based inference methods. This paper describes how survey data depart from the simple random sampling model implicit in most econometrics textbooks, points out where the benefits of accounting for this departure are likely to be greatest, and describes the relationship between design-based estimators and related econometric methods for regression.[1] In short, this paper is an economist's primer on survey samples.

A primer is needed because failure to account for how survey data are collected can, in some cases, lead standard econometric procedures to be seriously misleading. The effects of survey design are best illustrated with an empirical example. Table 1 presents a simple example of a human capital earnings function that relates individuals' earnings to factors such as education, age, sex, and race. Here we apply the model to data from the March 1997 Current Population Survey. The dependent variable is the log of hourly earnings, and the explanatory variables include years of schooling, age, and dummy variables for being black, Hispanic, or female. Column (1) of the table presents unweighted estimates of the model coefficients and standard errors based on the assumption of identically and independently distributed (IID) error terms. The results in column (1) indicate that blacks' wages are approximately 6% ($e^{-.063}$-1) lower than the wages of whites with similar characteristics. Although it includes fewer right-hand side variables than does most of the earnings regression literature, column (1) follows most of that literature in ignoring the sample design.

[1] Deaton (1997) provides an excellent review of some of the material covered in this article, with emphasis on design issues and techniques relevant to work using data collected in developing countries. The treatment here differs from Deaton's in being shorter and in emphasizing designs common to data produced by U.S. statistical organizations.

Column (2) of Table 1 presents estimates from the same model, but in that column the estimation procedure has explicitly accounted for the way in which the CPS data were collected. Some of the results are substantively unchanged, as the coefficients on education, age, and the female dummy change only slightly across the two specifications. The point estimate of the black/white wage differential changed substantially across the specifications, however, and each of the standard errors increased somewhat. The combined effect of the two changes was that the 95% confidence interval for the black coefficient is (-.090,-.036) for column 1 and (-.135, -.069) for column 2. The coefficient and confidence interval for the Hispanic dummy variable was similarly affected by the move to design-based estimation. These changes could obviously affect the results of statistical tests and other inference procedures.

Precisely what does it mean to "account for the survey design" in column (2)? Why did accounting for the survey design affect the inferences drawn from these data? Why were inferences for some parameters affected by the design while others were not? And finally, which set of estimates should an economist prefer? The answers to these and related questions are the subject of this paper. The paper describes design features common to surveys of households because most public-use data comes from such surveys, but also considers how surveys of businesses differ. The paper emphasizes linear regression, and it uses examples drawn from the National Longitudinal Survey of Youth of 1979 (NLSY79) and the Current Population Survey (CPS) to illustrate practical aspects of design-based estimation. We hope that the paper will enable the reader to be a more sophisticated consumer of survey data and, in turn, to make more accurate inferences from survey data.

2. Sample Design Basics

Survey samples are designed with specific goals in mind. The first goal of most surveys is to measure with reasonable accuracy the unconditional means or totals of key variables, or changes in these means or totals across time. For example, the monthly Current Population Survey (CPS) measures the national unemployment rate, and the Consumer Expenditure Survey (CEX) measures average spending patterns. The second goal of most surveys is to measure means for certain subgroups. For example, the

CPS measures the unemployment rate separately for each state, and the National Longitudinal Survey of Youth (NLSY) measures the labor market activity of black and white youth. These goals involve estimating means or related quantities that describe an existing *finite population* (for example, the current U.S. population). This is in contrast with a more typical goal for an economist: characterizing the underlying, often multivariate, behavioral relationship that generated the existing finite population. The divergent goals of survey designers and econometricians have led surveys to depart from the standard econometric sampling framework in several ways.

To fix ideas, suppose that a sample of size n of values of the variable Y is drawn from a finite population of size N. Typically n is much smaller than N and we will assume this to be the case throughout the paper. The most conceptually straightforward method of selecting a sample is *simple random sampling* (SRS), which is simply the urn model familiar from econometrics textbooks. SRS consists of a series of independent random draws, where each draw gives an equal chance of selection to all members of the population.[2] While SRS simplifies data analysis, it is rarely an optimal sample design given the survey goals described above. Instead, sample design trades off the costs of various design techniques against their effects on the precision of key statistics, leading to quite complex designs for most surveys.

While the details vary, most designs combine three basic features: *stratification*, *clustering*, and *varying probabilities of selection*. The following subsections describe these features and explain why they are used. Because design features' effects on simple statistics are most relevant to how and why they are used, the description in this section emphasizes estimators of means and totals. Section 3 gives a more detailed discussion of the effects of design on regression—a topic of central interest to economists, but usually a side issue to those designing (and funding) surveys.

2.1 Stratification

SRS may be an appropriate strategy when little is known about the population prior to the survey. In most cases, however, there is some *ex ante* information about the distribution of Y across subgroups or *strata* of the population identified by auxiliary variables. Stratification entails choosing independent subsamples of predetermined size from each stratum, thereby reducing sampling variation. The basic idea is that sampling variability of a sample mean can be divided into a) sampling variation within strata and b) sampling variation in each stratum's sample share. By fixing each stratum's sample share, stratification eliminates sampling variation due to this latter between-stratum component. If the strata have very different means for Y because the auxiliary variables used to define strata are highly correlated with Y, then stratification can increase precision substantially. If the strata are very similar to one another, however, then this procedure reduces sampling variance only slightly.[3] Thus, the ideal stratification scheme creates strata that are internally homogeneous and externally heterogeneous.

The effectiveness of stratification is limited by the type of information available prior to the survey. For example, a household income survey might wish to stratify on age, race, and sex because income is known to vary with these variables. There is no general list of households belonging to groups defined by those variables, however, so they cannot be used to directly define strata. To achieve similar results, sample designers often stratify the target population using averages for geographic areas to approximate the desired demographic stratification scheme. As an example, a survey for a city with equal shares of blacks and whites might sort neighborhoods into two strata based on whether the majority of residents are black or white. With a moderate degree of neighborhood segregation, the result might be one stratum with 70% black residents and the other with 70% white. The desired sample shares of each

---

[2] This describes SRS with replacement, but most surveys sample without replacement. Given that sample sizes are typically very small relative to the sizes of their target population, this distinction can usually be ignored. We will use SRS as a short hand for SRS with replacement.

[3] Often, different probabilities of selection are applied to elements in different strata, so stratification and varying probabilities of selection (the topic of section 2.2) are implemented together. Our description of the effects of stratification on precision applies when the same probability of selection is applied to each stratum. This isolates the

race can then be closely controlled through independent samples from each stratum. Though not as effective as direct stratification of households, such stratification of geographic areas is practical and widely used. Column 2 of Table 2 gives some examples. Samples for business surveys are usually selected from a list of establishments that has information on characteristics such as industry, location, and firm size for individual sampling units, so these characteristics are typically used as stratification variables. As Table 2 illustrates, business surveys are usually stratified by industry, and often by some measure of size as well.

2.2 Unequal probabilities of selection

An important feature of SRS is that it assigns the same probability of selection to each population element. Stratified samples can also use constant probabilities of selection—a technique known as proportional allocation. But in many cases varying the selection probabilities allows a survey to better achieve its goals. One reason for doing so is that it is easier to collect data from some members of the population than others. For example, face-to-face interviews are cheaper in the city than in the country.[4] Cost minimization subject to achieving a set level of precision will lead the sample designer to include urban households in the sample with higher probability than rural households.

A second reason for varying probabilities of selection is that a survey may need separate estimates for subpopulations of unequal size. For example, surveys such as the CPS that are designed to produce state estimates typically use higher rates of selection in states with small populations such as Alaska or Wyoming than in very populous states such as New York or California. A closely related third reason is that a survey's goal may be to compare subpopulations. If one group is smaller than the other, sampling the smaller group at a relatively high rate reduces the variance of the estimated difference. This

---

effect of choosing separate samples from each stratum (which defines stratification) from the effects of varying probabilities of selection.

[4] Deaton (1997, p. 12) points out that this is particularly true of surveys in third-world countries, where transportation and communication is much more costly in the hinterland than in the cities.

is one reason that many household surveys use higher rates of selection for black, Hispanic, or low-income families than for other families, as illustrated in the column (2) of Table 2.

A fourth reason for varying probabilities of selection is that some population elements are more informative than others. In business surveys in particular it is common to use *probability-proportional-to-size* sampling in which an element's probability of selection is set proportional to some measure of size such as a business's amount of sales or number of employees. This is illustrated in column 2, panel B of Table 2. Higher probabilities of selection reflect the greater importance of large businesses to aggregate measures such as GDP or to any variable that is measured per employee (e.g., rates of health insurance coverage).

Using estimators based on the assumption of SRS may result in biases if varying probabilities of selection are correlated with the target measure Y.[5] *Survey weights* are designed to allow the analyst to avoid such biases in making inferences about finite population parameters such as $\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$. Define π_i as the probability that element i is selected for the sample, which is a probability controlled by the sample designer. The basic sampling weight w_i is simply the inverse of this probability:

$$w_i = \frac{1}{\pi_i}. \qquad (1)$$

Often, w_i can be thought of as the number of finite population elements that sample observation i is intended to represent.

Equation (1) indicates that weights are based on parameters set by the sample design. In practice, however, survey weights are typically adjusted in two ways after the data have been collected. The first adjustment compensates for varying probabilities of response by shifting the weight of nonrespondents to observationally similar respondents.[6] Almost all surveys, even those designed to have equal probabilities

[5] For example, (unweighted) sample mean income from the NLSY, which sampled blacks, Hispanics, and low-income whites at much higher rates than other whites, is likely to underestimate true population mean income.

[6] *Item nonresponse* is a closely related problem that occurs when respondents answer some questions but not others. Surveys typically deal with this problem by "allocating" or "imputing" to nonrespondents the answers of otherwise-similar respondents who answered the particular question. For calculating means, this is a reasonable way of

of selection, make some such *nonresponse adjustment* to the survey weights. Nonresponse adjustments are implicitly based on a model of the determinants of response, and the model is necessarily limited to dependence on measures that are available for both nonrespondents and respondents. Column (3) of Table 2 lists the variables used in the nonresponse models of our example surveys.

The second adjustment to the survey weights, known as *post-stratification*, adjusts the weights to make weighted sample moments equal finite population moments that are known with a high degree of accuracy (perhaps from a much larger survey or from administrative records). For example, it is common for surveys of persons to adjust weights to make sample estimates match counts of the population by age, race, and sex from the decennial census of population. Column (4) of Table 2 lists the variables used for post-stratification in our example surveys. These variables are typically similar to the variables used to define the original stratification scheme. Post-stratification reduces the variance of estimates in much the way (pre-)stratification does; it is effective if the auxiliary variables used to post-stratify are highly correlated with the target variable.[7] An important difference between the two is that, under mild conditions, stratification at worst has no effect on precision, whereas post-stratification can reduce precision if it is based on variables not correlated with the variable of interest. This is of particular concern when using survey data for purposes very different from those that motivated the choice of poststratification variables.

2.3 Clustering

SRS involves independent selection of each sample element. However, collecting data from population elements that are close together is often considerably less costly than collecting data from elements chosen independently. In such cases, selecting groups of close elements (known as clusters)

---

making nonresponse adjustments on a question-by-question basis. Common imputation methods have much less innocuous effects on regression coefficients, however, which is one reason why analysts often delete observations with imputed data from their analysis. Most surveys in the U.S. have substantial rates of nonresponse for at least some questions. See Lillard, Smith, and Welch (1986) and Little and Rubin (1987) for further discussion.

[7] Post-stratification is often implemented using *raking*, which iteratively adjusts weights to reconcile post-stratification to moments of the marginal distributions of several variables. For example, a survey might use raking

reduces per-element collection costs, and thereby allows for a larger sample size than would SRS, holding costs fixed. This is particularly true in face-to-face surveys for which interviewer travel time accounts for a large share of collection costs, and so face-to-face surveys of households virtually always use clustered samples. Column (5) of Table 2 lists the variables used to define cluster in our example surveys. For household surveys, counties, groups of counties, or MSAs are most often used at the initial stage of clustering, and blocks or groups of blocks are often used at the second stage of clustering. Business surveys are rarely explicitly clustered, but analysis of elements at levels below the business as a whole (e.g., business units, plants, or employees) are implicitly clustered, since employees from a single business were selected for the sample at the same time.

Although clustering reduces survey costs per element, the value of a target variable may be correlated across elements belonging to the same cluster. For example, the incomes of two randomly selected families from the same neighborhood are more alike on average than the incomes of two families selected independently from the U.S. as a whole. Such within-cluster correlation reduces the precision of estimators relative to what could be obtained from an SRS sample of the same size because two selections from the same cluster provide less information than two independent selections. If clusters are composed of identical elements—in which case selecting more than one element from a cluster yields no additional information—then the sample size is effectively the number of clusters and collecting data from more than one cluster member simply wastes resources. If instead each cluster is effectively a simple random sample of the population at large, then a clustered sample would be just as informative as a simple random sample of the same size.

While data collection methods determine the relationship between survey costs and the geographic dispersion of sample elements, defining clusters is part of the sample design. The ideal clustering scheme creates clusters that are spatially compact (thereby minimizing survey collection costs)

---

to post-stratify to population by age group and population by level of education (when population by age group by level of education is not known with great precision).

but internally heterogeneous (thereby maximizing the information captured in each element).[8] A fairly typical design for a large U.S. household survey would use two levels of clustering: divide the area of the U.S. into counties or small groups of counties and choose a sample of those clusters;[9] then subdivide each sampled area into much smaller units (perhaps a city block), and choose a sample of those units; individual elements (households) are then selected randomly from the second-stage units. Counties are often chosen as first-stage clusters because for many variables they are internally fairly heterogeneous while being compact enough to realize much of the cost savings from reducing the geographic dispersion of data collection.

3. Regression Coefficients and Survey Weights

The previous section described the components of sample design in terms of the estimation problems that typically motivate the choice of design—estimation of means or totals. Yet economists more often use survey data for multivariate analyses, so the effects of sample design on regression and related estimators are of more fundamental concern for economic analyses. This section examines the consequences of varying probabilities of selection for the OLS coefficient estimator, and considers when using survey weights will improve its properties. The consequences of stratified and clustered samples are considered in sections 4 and 5, where we take up the issue of variance estimation.

3.1 *Differences in Approach*

Before considering the effects of sample design it is helpful to point out differences between the finite population approach to regression taken by survey statisticians and the modeling approach that underlies most econometric analyses. To understand the different approaches, it is important to keep in

[8] Note the distinction here between clusters and strata. Because stratification ensures that each stratum is represented in the survey, its benefits are greatest when the strata are quite different from one another, and so are internally relatively homogeneous. In contrast, it is advantageous that clusters be internally heterogeneous, and consequently that different clusters within a stratum be somewhat similar to each other. If this is not true, then the fact that clusters are selected rather than individual elements will substantially increase variance.

[9] These first-stage clusters are often referred to as primary sampling units (PSUs).

mind that there are two random processes involved in economic survey data. First, a data-generating process produces population elements such as individuals, firms, or countries in a manner that may or may not be independent and identically distributed (IID). These population elements constitute the finite population. Second, a sampling process produces the sample from the finite population. The manner in which finite population elements are selected for the sample may or may not be SRS. A key difference between the two disciplines is that econometricians focus on the vagaries of the data-generating process, while survey statisticians focus on the sampling process.

Consider the multiple regression model:

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{e} \tag{2}$$

where $\mathbf{X}$ has full column rank k, $E(\mathbf{e}|\mathbf{X})=0$, and $E(\mathbf{åå}'|\mathbf{X}) = \mathbf{ó}^2 I$. We will think of (2) as a data-generating process that could potentially produce an infinite number of observations. The finite population is viewed as a sample of size N produced by this process, and the survey sample is in turn drawn from the finite population. If SRS is used to select a survey sample, then the properties of the data-generating process also apply to the sampled data. But under more complex designs the joint distribution of sample values becomes more complicated. We will use lower-case letters to denote sample values, upper-case letters to refer to random variables more generally, and the subscript N to denote finite population values.

The classical statistics problem is to make inferences about β. In contrast, the sampling literature frequently defines the parameter of interest as the finite-population quantity:

$$B = (X_N' X_N)^{-1}(X_N' Y_N). \tag{3}$$

That is, the quantity of interest is the vector that would be obtained by applying least squares to the entire finite population. Note that (3) exists whether or not the finite population is generated by the mechanism in (2), though its interpretation and interest may depend on the accuracy of the model.

If B is the parameter of interest, then an estimator's properties are evaluated by taking expectations over all possible samples that could be drawn from the finite population, with relevant

probabilities determined by the sample design. We will refer to this as taking expectations with respect to the design ($E_{Design}\{ . \}$). Economists would more typically evaluate the properties of an estimator of β (rather than B) by taking expectations over all possible outcomes of a particular data-generating process, under the assumption that each sample element is a random draw from that process. We will refer to this as taking expectations with respect to the model ($E_{Model}\{ . \}$).

Note that taking expectations with respect to the design yields a function of finite population quantities that can be thought of as outcomes of the data-generating mechanism. These two approaches can be combined by taking expectations first with respect to the design (but assuming that the finite population was generated by a process described by the model), and then with respect to the model—that is $E_{Model}(E_{Design}(. |\text{model assumptions}))$. Doing so allows consideration of how the sample design affects estimator behavior under a particular set of assumptions about the data generating mechanism.

3.2 OLS with Unequal Selection Probabilities

Under SRS, the classical estimator of both β in (2) and B in (3) is the ordinary least squares (OLS) estimator

$$\hat{b}_{OLS} = (x'x)^{-1}x'y, \qquad (4)$$

where lower-case letters denote sample values. For samples with varying probabilities of selection, survey statisticians would typically recommend use of the weighted least squares (WLS) estimator:

$$\hat{\boldsymbol{b}}_{wgt} = (x'wx)^{-1}x'wy \qquad (5)$$

where *w* is a diagonal matrix of survey weights. What should guide an economist's choice between the two estimators? Note that the econometric argument for WLS is usually based on heteroscedasticity which is assumed away in (2), so any preference for WLS must be driven by other factors. Here, instead, the argument for WLS is that varying probabilities of selection may lead the relationship between the dependent variable and regressors in the sampling distribution to differ from the relationship in the finite

population. If the sample and population relationships differ, the consistency of OLS will depend on how the sample was selected. Because it uses sample weights that are inversely proportional to a sample element's probability of selection, $\hat{b}_{WGT}$ converges to the finite population relationship B even in cases where $\hat{b}_{OLS}$ does not. The rest of this subsection is devoted to making this point more precisely.

Taking expectations with respect to the design under (2), gives

$$E_{Design}\{\hat{b}_{OLS} - b\} = E_{Design}\{(x'x)^{-1}x'e)\} \xrightarrow[n\to\infty]{} (X'_N p_N X_N)^{-1} X'_N p_N e_N, \quad (6)$$

where p_N denotes a diagonal matrix with probabilities of selection along the diagonal, and e_N is the finite-population matrix of error terms. Taking the expectation of (6) with respect to the model, $E_{Model}\{(X'_N p_N X_N)^{-1} X'_N p_N e_N\}$ will converge to zero provided that $\lim_{N\to\infty} (X'_N p_N X_N)^{-1}$ exists, and $E_{Model}(N^{-1}\sum_{i=1}^{N} X_i p_i e_i) \xrightarrow[N\to\infty]{} 0$. This latter condition will hold if the data-generating process is such that the product of the sampling probability and the error term is on average zero for every value of X: $E_{Model}(pe \mid X) = 0$. With constant probabilities of selection, this reduces to the more familiar condition $E_{Model}(\mathbf{e}|\mathbf{X})=0$. However, this condition is not sufficient to ensure that the OLS estimator is consistent when probabilities of selection vary.

If probabilities of selection (π) are correlated with $\mathbf{e}$, then the mean of the error term's sampling distribution will not equal zero for all X. As should be familiar from the econometric literature, this leads to inconsistent slope coefficient estimates if the mean of $\mathbf{e}$ (in the sampling distribution) is correlated with X. Thus, consistency of OLS applied to a complex sample not only requires that (2) be correctly specified, but also that the product $\pi\mathbf{e}$ be uncorrelated with X.

When is this second condition likely to hold? For many sample designs, probabilities of selection can be expressed as a linear function of design variables D: $\pi_i = d_i\tau$.[10] For example, D includes stratum

[10] π_i will be a linear function of design variables if sampling probabilities are constant within a stratum, or are proportional to some measure of size. Note, however, that probabilities of response may be non-linear functions of these variables or functions of variables not included in D. Other factors, such as imperfections in the information

identifiers if probabilities of selection vary across strata, and includes measures of size if probability-proportional-to-size sampling is used. In some cases, D also includes Y. For example, a survey might oversample low-income families, in which case the dependent variable from an earnings regression would also be a design variable.

If X includes all columns of D, then π is constant conditional on X, and there is arguably no reason to prefer $\hat{\boldsymbol{b}}_{wgt}$ to $\hat{\boldsymbol{b}}_{OLS}$. If, in addition, (2) is known to be correctly specified, then $\hat{\boldsymbol{b}}_{OLS}$ is the minimum-variance linear unbiased estimator and so would be preferred to $\hat{\boldsymbol{b}}_{wgt}$. However, only rarely would all elements of D be included in X. For example, in many household surveys probabilities of selection vary across geographic areas. Geographic identifiers may not be included in X either because doing so would be inappropriate for the analysis, or because the relevant geographic detail is not identified in public use files. In cases in which D includes the model's dependent variable, clearly it is not possible to simply include D as a regressor.[11] Thus, the conditions under which $\hat{\boldsymbol{b}}_{OLS}$ is strictly preferred to $\hat{\boldsymbol{b}}_{wgt}$ are not often met when using survey data.

3.3 *Advantages of Weighting*

If X does not include all of D, and so π_i varies across members of the population with the same set of X's, then $\pi\mathbf{e}$ could be correlated with X. Consistency of $\hat{\boldsymbol{b}}_{OLS}$ then rests on the *assumption* that $\pi\mathbf{e}$ is uncorrelated with X.[12] If there were great certainty about the adequacy of the model—that is, that conditional on X all variation in Y were purely the result of white noise shocks—then there would be little reason to suppose that the selection probabilities and the error term were correlated. But more

---

used to design the sample, may also mean that simply including design variables as regressors will not guarantee consistency of OLS.

[11] In this case the sample design produces a form of selection bias. Hausman and Wise (1981) and Wooldridge (1998, 1999) consider this case.

[12] A more intuitive, but more restrictive condition, is that $E_{Model}(e \mid X, p) = 0$, or, yet more restrictive, $E_{Model}(e \mid X, D) = 0$.

realistically there is always some uncertainty about the specification, in which case $\hat{\boldsymbol{b}}_{wgt}$ may be preferred because it has a smaller bias than $\hat{\boldsymbol{b}}_{OLS}$ when the model is misspecified.

The argument that $\hat{\boldsymbol{b}}_{wgt}$ has advantages over $\hat{\boldsymbol{b}}_{OLS}$ is best illustrated using a particular type of misspecification. Suppose that (2) omits relevant variables **Z,** so the correctly specified model is:

$$\mathbf{Y} = \mathbf{X}\beta_z + \mathbf{Z}\gamma + \nu. \qquad (6)$$

Neither OLS nor WLS will be consistent for $\boldsymbol{b}_Z$ when applied to (2), except under restrictive conditions. Obviously the ideal solution would be to include *z* in the regression, but we will assume that *z* is unavailable, as is often the case in economic survey data.

Since we cannot estimate $\boldsymbol{b}_Z$, suppose that instead we take as our parameter of interest $\boldsymbol{b}^* = \boldsymbol{b}_Z + T^*\boldsymbol{g}$, where $T^* = \lim_{N\to\infty}\{(X_N' X_N)^{-1} X_N' Z_N\}$. That is, we cannot avoid attributing some of the effects of Z to X, but we want our estimates to reflect how those variables relate to each other under the mechanism that generated the finite population.[13] Kott (1991) shows that $E_{Model}\{E_{Design}\{\hat{\boldsymbol{b}}_{OLS}\}\}$ converges to $\boldsymbol{b}^*$ only if p is uncorrelated with the components of Z that are orthogonal to X. In contrast, $E_{Model}\{E_{Design}\{\hat{b}_{wgt}\}\}$ converges to $\boldsymbol{b}^*$ regardless of the structure of the p's. In this sense, $\hat{\boldsymbol{b}}_{wgt}$ is more robust than $\hat{\boldsymbol{b}}_{OLS}$. Thus the prime argument for using $\hat{\boldsymbol{b}}_{wgt}$ is that it consistently estimates a well-defined quantity (B or $\boldsymbol{b}^*$) even when the model is misspecified, whereas $\hat{\boldsymbol{b}}_{OLS}$ estimates a well-defined quantity only under more restrictive conditions. The primary cost of using $\hat{\boldsymbol{b}}_{wgt}$ is that it has greater variance than $\hat{\boldsymbol{b}}_{OLS}$. In general, $V(\hat{b}_{wgt})$ increases as the variance of weights increases and as the sample size decreases (Pfefferman 1996).[14]

[13] This sort of compromise is prevalent in the econometrics literature on returns to education (e.g., Freeman, 1986), to take just one example.

[14] A secondary cost is that some standard inference procedures are unavailable when using the weighted estimator—for example likelihood ratio tests and use of residuals to check the fit of a model (Pfeffermann 1996)

The effects of weighting when controlling for design variables are illustrated in Table 3, which revisits the human capital earnings function of Table 1. The first two columns of Table 3 reproduce coefficients of Table 1, and they again show that weighting affects some variables but not others. In particular, the negative coefficients on the black and Hispanic dummy variables are significantly larger in absolute value when the weighted estimates are used, whereas the coefficients on age, years of schooling, and the female dummy are largely unaffected by the weighting. The CPS sample design is the cause of this pattern. In particular, because it is used to estimate state unemployment rates, the CPS design selects households in small states with higher probabilities than households in large states. Many of these less populous states are primarily non-urban, while also having populations that are disproportionately non-black and non-Hispanic and which have lower-than-average earnings for given characteristics (i.e. $\bar{e} < 0$ for many of these states). Thus, the unweighted CPS tends to overrepresent low-earning whites, but not low-earning blacks or Hispanics. As a result, the unweighted black-white and Hispanic-non-Hispanic comparisons of column (1) are inaccurately close to zero. In contrast, the overrepresentation of small states in the CPS has only modest effects on the coefficients on age, schooling, and sex. The reason is that there is much less cross-state variation in the population along these dimensions (although there is some).

Columns (3) and (4) of Table 3 reproduce the analyses of the first two columns, with the exception that state dummy variables are included as regressors. Recall again that most of the cross-household variation in the probability of selection in the CPS is across state. Thus, the state dummies are the primary design variables in D.[15] The columns show that the effect of weighting is much smaller when these design variables are included on the right-hand side. This is not to suggest that design variables should always be included as right-hand side variables, as in many contexts (perhaps including this one) the relationships of interest do not warrant their inclusion. Rather, the point is merely that the effects of

[15] State dummies explain 87% of the variation in the sampling weights in the CPS, where the "sampling weight" is the weight without any nonresponse or poststratification adjustment. State dummies explain 61% of the variation in the survey weight for the entire sample of persons, and 63% of the variation in the survey weight for our regression sample.

weighting depend on whether or not the design variables are included as regressors.

The argument that $\hat{\boldsymbol{b}}_{wgt}$ is more robust than $\hat{\boldsymbol{b}}_{OLS}$ assumes that the variation in survey weights accurately reflects differences in probabilities of sample inclusion. In practice, the weights provided may do so imperfectly. For example, nonresponse adjustment is necessarily based on a model, and that model may itself be misspecified. If so, adjusting weights for nonresponse is likely to increase estimator variances, and there could be some analyses in which using weights would increase bias rather than reduce it. The need to adjust weights for missing data also arises when analysts impose selection criteria of their own—for example dropping observations with missing responses to relevant questions. The correct sample weights would account for this stage of selection as well, and for some analyses the correct weights might not be highly correlated with the weights constructed for the survey sample as a whole.[16] Note that these may be strong arguments for tailoring selection adjustments to a particular analysis, but not for choosing to ignore the problem.[17]

3.4 *Using Weights to Check for Misspecification*

Even in cases where an economist is doubtful that $\hat{\boldsymbol{b}}_{wgt}$ has properties superior to those of $\hat{\boldsymbol{b}}_{OLS}$, the weights themselves may be useful in checking for possible misspecification. Weighted estimates differ substantially from unweighted estimates when there are large between-group differences in both probabilities of selection and in the relationships represented by the regression coefficients. With a correctly specified model (i.e. one that allows for between-group differences in regression coefficients where relationships differ across groups), there should be little difference between the weighted and unweighted coefficient estimates. When a model fails this check, differences between the two sets of estimates are likely due to misspecification that is related to the variables that determine probabilities of selection. Restricting one's attention to unweighted estimators thus ignores a simple and potentially

[16] This can be particularly problematic in longitudinal surveys (MaCurdy, Mroz and Gritz, 1998).

important check on a model's robustness.

DuMouchel and Duncan (1983) propose one method of testing for misspecification using survey weights. They suggest adding to the list of regressors the weights (w) along with interactions between w and each variable in x, and then testing for the joint significance of the coefficients on w and its interactions. Rejecting the null that those coefficients are jointly zero is evidence that the model may be misspecified. A statistically significant interaction between w and a component of x may indicate that the effects of that component are misspecified. So, using DuMouchel and Duncan's example, if a survey uses higher sampling probabilities for blacks than whites, and the coefficient on an interaction between w and schooling attainment is significant, then this may indicate that the schooling coefficient differs by race.[18]

4. Effects of Sample Design on Conventional Variance Estimators

Ignoring varying probabilities of selection may lead to biased coefficient estimates. Stratification and clustering, in contrast, do not affect the means of point estimators but can have important effects on their variances. In this section, we consider how these techniques affect the consistency of variance estimators that assume SRS. In section 5 we present alternative methods of variance estimation that take the sample design into account.

4.1 Stratification

We noted in section 2 that stratification eliminates the contribution of between-stratum differences to the variability of estimators of means or totals. The corresponding result for regression is that stratification may reduce $V(\hat{b})$ by eliminating the contribution of between-stratum differences in the

[17] One alternative might be use of Hellerstein and Imbens's (1999) method of constructing weights for a particular regression analysis using moment restrictions estimated from auxiliary data—essentially a regression-based form of post-stratification.

[18] In a similar vein, Wooldridge (1998,1999) derives a Hausman test based on the difference between weighted and unweighted estimators which tests for the exogeneity of the sampling probabilities. See also Fuller (1984) and Pfefferman (1993, 1996) for further discussion.

mean of the error term. As a result, ignoring stratification leads, on average, to overestimates of $V(\hat{\beta})$. Recall, however, that the limited information available ex ante for most surveys means that gains from stratification tend to be small even in estimating $\overline{Y}$. With regression, variation in X across strata usually accounts for some of the variation in Y, so differences across strata in mean ε will generally be even smaller than differences in mean Y. Thus, stratification leads to efficiency gains that are typically quite small in the regression context, and for this reason, ignoring stratification in the estimation of variances is not likely to cause substantial biases. Nevertheless, accounting for stratification will on average lead to lower variance estimates and so is advisable when estimation techniques that account for stratification and stratum identifiers are readily available,

4.2 Clustering

Clustering can lead to a violation of the assumption that error terms are independently distributed. Two elements selected within the same cluster will be more alike along many dimensions than would two elements selected independently from the population at large. As is familiar to economists from other contexts (e.g., autocorrelated errors, panel data models) positively correlated error terms result in a downward bias in conventional estimators of standard errors. Thus, ignoring clustering can lead to seriously misleading inference procedures. The magnitude of this downward bias depends on several factors: the correlation between error terms within clusters, the number of elements selected within a cluster, and how much the explanatory variables vary within clusters.

The effects of clustering can be characterized in terms of the ratio of a) the true variance of an estimator (taking the clustering into account) to b) the expectation of the variance estimator actually used. This quantity is known as a misspecification effect (*meff*), and it measures the degree to which variances are understated if clustering is ignored.[19] To characterize the magnitude of meffs in clustered data, suppose that a design involves choosing a SRS of α clusters, and then within each cluster choosing a SRS

of $\boldsymbol{b}_c$ elements. The formulae for meffs are most readily interpreted in the special case where the correlation among observations in a cluster can be modeled as due to a cluster effect in the error term. That is, the error term in $\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}$ may be modeled as the sum of two components: $\boldsymbol{e}_{ic} = u_c + \boldsymbol{n}_{ic}$, where $\text{cov}(u_c, \boldsymbol{n}_{ic}) = 0$. Treating the cluster effects (u_c) as random implies that:

$$V(\boldsymbol{e}_{ic}) = \sigma_u^2 + \sigma_{\boldsymbol{n}}^2, \quad \text{and} \quad \rho_{\boldsymbol{e}} = Corr(\boldsymbol{e}_{ic}, \boldsymbol{e}_{i'c}) = \sigma_u^2 \Big/ (\sigma_u^2 + \sigma_{\boldsymbol{n}}^2). \tag{7}$$

These formulae bear an obvious resemblance to random effects models used for panel data, a similarity we return to in Section 5. [20]

Scott and Holt (1982) show that in the case of random cluster effects, the ratio of $V(\hat{\boldsymbol{b}})$ to $\sigma^2 (X'X)^{-1}$ (which is a matrix of meffs) is given by

$$M = I + (\Lambda - I)\rho_e, \tag{8}$$

where $\Lambda = \left[\sum_{c=1}^{a} b_c \bar{X}_c' \bar{X}_c\right](X'X)^{-1}$, $\bar{X}_c = \left[\frac{1}{b_c} 1_{b_c} 1_{b_c}'\right] X_c$, X_c is the matrix of explanatory variables for cluster c, and 1_{b_c} is a row vector of 1's of length b_c. Here $\bar{X}_c$ is a $\boldsymbol{b}_c$ by $\boldsymbol{k}$ matrix of cluster means, and Λ is thus a function of across-cluster variation in X relative to total variation in X. In the special case of a regression with one regressor and with the same number of elements sampled from each cluster ($\boldsymbol{b}_c = \boldsymbol{b}$), the meff for the slope coefficient variance further simplifies to:

$$meff(\hat{b}) = 1 + (b-1)\hat{\rho}_x \rho_e \tag{9}$$

where $\hat{\rho}_x$ is the sample intra-cluster correlation of X. [21]

---

[19] Meffs are a general tool used to characterize biases in estimating variances for complex sample designs—they are not specific to designs with clustering.

[20] Indeed, models of this form have been considered in both the sampling literature (Kott 1991) and the econometrics literature (Moulton 1986, 1990).

[21] That is, $\hat{\rho}_x = (b\lambda - 1)/(b-1)$, and $\lambda = b \sum_{c=1}^{a} (\bar{X}_c - \bar{X})^2 \Big/ \sum_{c=1}^{a} \sum_{i=1}^{b} (X_{ci} - \bar{X})^2 \leq 1$ is the ratio of the between-cluster sum of squares to the total sum of squares (a univariate version of Λ).

The random effects model in (7) makes quite restrictive assumptions about the error structure, and so meffs based on it will be accurate only when the assumptions hold. Nevertheless, (8) and (9) provide a useful indication of where the variance estimator biases from ignoring clustering are likely to be most severe. For example, conventionally estimated standard errors for the coefficients of geographic controls (which usually do not vary at all within cluster and so have $\hat{r}_x = 1$) will generally have greater downward biases than will standard errors for demographic variables such as age. In addition, demographic characteristics with a high degree of geographic segregation (for example, race) generally have larger meffs than characteristics more evenly distributed across areas (for example, gender).

Equation (9) also illustrates that ignoring the effects of clustering generally leads to larger downward biases for the variances of means than for the variances of regression coefficients. With no explanatory variables, (9) simplifies to $meff(\hat{b}_0) = meff(\bar{y}) = 1 + (b-1)r_y$ where r_y is the intra-cluster correlation coefficient for Y. In the worst-case scenario for a regression (no within-cluster variation in X), $\Lambda = bI$, and conventional variance estimators are biased by a factor $1 + (b-1)r_e$. This expression differs from the meff for the sample mean only in that the correlation coefficient is for the error term rather than the dependent variable itself. Meffs for regression estimates tend to be smaller than meffs for sample means for two reasons: (i) the within-cluster correlation of residuals will generally be smaller than the correlation for the dependent variable itself ($r_e < r_y$) because regressors usually control for some within-cluster correlation; (ii) within-cluster variation in the regressors ($\hat{r}_x < 1$) also means that additional observations from a given cluster are informative even when the error terms are highly correlated.

Table 4 illustrates the effects of clustering on the estimation of human capital earnings functions with data drawn from the CPS and the NLSY79. Column (1) reports the univariate intracluster correlation coefficient for each of the explanatory variables in the regression.[22] The column shows that there is

[22] For both surveys, the counties (or groups of counties) selected as primary sampling units are used as clusters. Both samples have a second stage of clustering, which we ignore here but return to later. For both surveys, the

substantial variation in the extent to which independent variables are correlated within clusters. In household surveys, there is often little within-cluster correlation in variables such as age and sex, as these variables are spread about as evenly across primary sampling units as would be predicted by random assignment. In contrast, variables measured at aggregate levels such as the local unemployment rate or a state identifier will often be perfectly correlated within clusters. In between these two extremes lie variables such as education or race for which there is substantial but incomplete spatial segregation, and thus middling levels of intracluster correlation.

The intracluster correlation coefficient of the regression residual ($\hat{r}_e$) is reported above the list of explanatory variables for each survey. In contrast to the intracluster correlation coefficient for the right-hand side variables, this correlation is constant across all variables. Though $\hat{r}_e$ is small for both surveys, the effects of clustering may not be small because both surveys have a fairly large number of observations per cluster (on average 47 persons in the CPS regression sample, and 45 persons in the NLSY79 sample). Columns 2 and 3 report the square roots of two alternative meffs. (The square root is known as a misspecification factor, or *meft*, which gives the effects of survey design on standard errors rather than variances.) The last column indicates that accurately-estimated standard errors sometimes substantially exceed those calculated assuming IID errors. For example, the misspecification factor for the MSA dummy coefficient in the CPS regression is estimated to be 1.24, which suggests that inference procedures based on conventional standard errors could be quite misleading.

Note that equation (9), which is based on the random effects model, suggests that the misspecification effect should be a monotonic function of the intracluster correlation coefficient for independent variables in the same regression. There are two reasons why this monotonicity is not observed in the table. First, column 1 reports the univariate intracluster correlation coefficient for each X,

---

sample design involved selection of one PSU per non-certainty stratum. Design-based variance estimators require a minimum of two PSUs per stratum, so non-certainty strata were paired to approximate such a design. The estimates in Table 4 are all unweighted, to facilitate use of the random components model.

whereas in the multivariate framework the relevant calculations involve the full matrix Λ.[23] This accounts for the difference between the results in column 2, and what would be implied by applying equation (9). Second, the reported misspecification effects are based on techniques (discussed in section 5) that allow for a more general error structure than does the random effects model of (7). That is, while (9) provides useful intuition, it will be accurate only in special cases.

Treating clustered samples as if each observation were drawn independently has much the same effect as treating panel data as if it were a series of independent cross sections: in most cases the variance estimator will have a downward bias. With panel data, standard practice among economists is now to account for the potential correlation of error terms in estimation. It ought also be standard practice to account for correlation induced by the sample design. The next section discusses methods for doing so, and the practical difficulties in implementing these methods.

5. Variance estimators for data from complex samples

With data from a clustered sample, the variance-covariance matrix of the regression error (Ω) does not in general equal σ^2I. This implies that in the presence of clustering OLS is not efficient and that variance estimators based on the assumption of SRS are not consistent. Economists might approach this problem by assuming a particular form for Ω (such as the random effects model of (7)) and then using that assumption as the basis for both a more efficient estimator of β and a consistent estimator of the associated variance-covariance matrix. If the implied restrictions on Ω are valid, then this approach is both unbiased and efficient. Survey statisticians, in contrast to economists, are generally unwilling to impose additional structure on Ω. Instead, they typically choose to forego potential efficiency gains in

[23] Variation in cluster size plays a role as well, because cluster size is in some cases correlated with the explanatory variables and as a result it affects some variances more than others.

exchange for the greater robustness of $\hat{b}_{wgt}$, and so the sampling literature concentrates instead on developing estimators of V($\hat{b}_{wgt}$) that account for survey sampling methods.[24]

This section of the paper describes the estimators of V($\hat{b}_{wgt}$) devised by survey statisticians, and then compares them with more traditional econometric variance estimators. From a repeated sampling perspective, $\hat{b}_{wgt}$ is a non-linear estimator as both X and Y vary across samples. The survey statistics literature has developed two general approaches to estimating the variance of a nonlinear estimator: "linearization" estimators, based on the variance of a linear approximation to the nonlinear estimator; and replication estimators, based on variation in estimates across repeated subsamples of the survey data. While methods of computation are quite different for the two types, in practice they have been found to have similar properties, so the key issue in choosing between them is generally convenience rather than their statistical properties. In what follows, we describe these approaches and consider how they relate to methods based on a random effects model.

5.1 Linearization Estimators

Linearization estimators use the first term of a Taylor expansion to approximate a nonlinear estimator with a linear function of estimated (finite population) totals. For example, the weighted sample mean ($\bar{y}_{wgt} = \sum w_i y_i / \sum w_i = \hat{Y}_N / \hat{N}$), can be expanded around the true finite population values to obtain an approximation in terms of the totals Y_N and N:

$$\bar{y}_{wgt} \approx \frac{Y_N}{N} + \frac{1}{N}(\hat{Y}_N - Y_N) - \frac{Y_N}{N^2}(\hat{N} - N).$$

This approach breaks the problem into two steps: first estimate the coefficients in the linear approximation, and then estimate the variance of the totals. Techniques developed to estimate the variance of a total with data from complex samples thus also form the basis for linearization estimators

[24] We present methods for the weighted least squares estimator. Results for the OLS estimator are simply derived by treating the weights as all equal to 1, but are subject to the point estimation bias issues raised in section 3.

for variances of regression coefficients. As a result, a brief introduction to those techniques provides a useful starting point.

5.1.1 Variance Estimators for Finite Population Totals

Consider estimation of the finite population total $Y_N = \sum_{i=1}^{N} Y_i$ using data from a stratified clustered sample with unequal probabilities of selection—i.e. a sample that combines the design elements described in section 2. Let H denote the number of strata, and note that the population total is simply the sum of the stratum totals, so $\hat{Y}_N = \sum_{h=1}^{H} \hat{Y}_h$ where $\hat{Y}_h$ denotes the estimated total for stratum h. Stratification involves choosing an independent sample from each stratum, so $V(\hat{Y}_N) = \sum_{h=1}^{H} V(\hat{Y}_h)$. Thus, stratification is handled by estimating variances separately within each stratum and then summing. To simplify discussion of estimators of $V(\hat{Y}_N)$ we present estimators that would apply to data from a single stratum.

To make the intuition clearer, think of decomposing an element's weight, w_{ci}, as $w_{ci} = w_c \times w_{i|c}$, where w_c gives the inverse of the probability that cluster *c* is selected, and $w_{i|c}$ gives the inverse of the probability that element *i* is selected given that cluster *c* is in the sample. Let $\mathbf{B_c}$ denote the total number of elements in cluster c, while $\mathbf{b_c}$ continues to denote the number selected. Similarly, **A** is the total number of clusters, and **a** the number selected. The estimator for the total would then be $\hat{Y}_N = \sum_{c=1}^{a} w_{ci} y_{ci} = \sum_{c=1}^{a} w_c (\sum_{i=1}^{b_c} w_{i|c} y_{ci}) = \sum_{c=1}^{a} w_c \hat{Y}_c$, where $\hat{Y}_c$ is an estimate of the total for sample cluster *c* ($Y_c = \sum_{i=1}^{B_c} y_{ci}$). Suppose that the design involves a SRS of **a** clusters, and that all elements within sample clusters are included in the sample ($w_{i|c}$ = 1, $b_c = B_c$). In this case, Y_c is known with certainty for each cluster in the sample. The variance of the total is then estimated as:

$$\hat{V}(\hat{Y}_N) = \frac{a}{a-1}\sum_{c=1}^{a}(w_c Y_c - \overline{w_c Y_c})^2, \quad \text{where} \quad \overline{w_c Y_c} = \frac{1}{a}\sum_{c=1}^{a} w_c Y_c \,.^{25} \tag{10}$$

If there is subsampling within clusters ($w_{i|hc} > 1$) the variance of the total is commonly estimated using:

$$\hat{V}(\hat{Y}_N) = \frac{a}{a-1}\sum_{c=1}^{a}\left(w_c \hat{Y}_c - \overline{w_c \hat{Y}_c}\right)^2 .^{26} \tag{11}$$

The similarity of (10) and (11) suggests that use of the latter ignores the effects of subsampling (and, consequently, of any additional stages of clustering). However, in expectation (11) exceeds (10) by a term that roughly equals the contribution of the within-cluster sampling design to V($\hat{Y}_N$).[27] Sampling variation in $\hat{Y}_c$ means that substitution of $\hat{Y}_c$ for Y_c increases the average size of the term in parentheses; the size of that increase depends on the variance of $\hat{Y}_c$ as an estimator of Y_c; and the variance of $\hat{Y}_c$ depends on the within-cluster sampling design.

5.1.2 Linearization Variance Estimators for Regression Coefficients

Returning to regression, we want to approximate $\hat{b}_{wgt}$ with a linear function of estimated (finite population) totals to develop a linearization variance estimator for $\hat{b}_{wgt}$. To this end, consider the parameter vector B defined implicitly (as in equation (3)) by

$$X_N'(Y_N - X_N B) = \sum_{i=1}^{N} X_i' e_i = 0 . \tag{12}$$

---

[25] This formula may seem more intuitive if the reader notes that when clusters are selected with SRS $w_c = \frac{A}{a}$ and $\hat{Y}_N = A\overline{Y_c}$. Then (10) simplifies to $\hat{V}(\hat{Y}_N) = A^2\hat{V}(\overline{Y_c}) = A^2\frac{1}{a}s^2_{Y_c}$, where $s^2_{Y_c} = \frac{1}{a-1}\sum_{c=1}^{a}\left(Y_c - \overline{Y_c}\right)^2$.

[26] This is an approximately unbiased estimator when first-stage sampling units (clusters) are chosen with replacement using probability-proportional-to-size sampling. It is more common to sample without replacement (because it makes estimators more efficient), but estimating variances is much simpler if samples are treated as if they were chosen with replacement. This leads to an upward bias in estimating the variance, but the relative bias will generally be small if the first stage sampling fraction is small. See Rao, Wu, and Yue (1992).

[27] See, for example, Wolter (1985, p. 46) and Shao (1996).

That is, B is defined by the least squares normal equations applied to the entire finite population.[28]

Denote the left hand side of (12) as U_N and note that it is a vector of finite population totals

($U_N = \sum_{i=1}^{N} X_i' e_i$). So if e were observable (and hence B were known) we could estimate U_N using the

sample total $\hat{U}_N = \sum_{c=1}^{a}\sum_{i=1}^{b_c} w_{ci} x_{ci}' e_{ci} = x'we$. $\hat{U}_N$ is assumed to be asymptotically normal with mean zero

and variance-covariance matrix Σ_U.

In addition, note that the estimator $\hat{b}_{wgt}$ is the solution of the analogous estimating equation

$\sum_{c=1}^{a}\sum_{i=1}^{b_c} w_{ci} x_{ci}' (y_{ci} - \hat{b}_{wgt} x_{ci}) = x'we = 0,$ where e is simply the vector of residuals. Let $\hat{U}_N\left(\hat{b}_{wgt}\right) = x'we$.

Using the first term of a Taylor expansion of $\hat{U}_N\left(\hat{b}_{wgt}\right)$ around $\hat{b}_{wgt} = \mathrm{B}$ as an approximation:

$$0 = \hat{U}_N(\hat{b}_{wgt}) \approx \hat{U}_N(\mathrm{B}) + \frac{\partial \hat{U}_N(\mathrm{B})}{\partial \mathrm{B}}(\hat{b}_{wgt} - \mathrm{B}) \tag{13}$$

which gives $\hat{U}_N(\mathrm{B}) \approx -\frac{\partial \hat{U}_N(\mathrm{B})}{\partial \mathrm{B}}(\hat{b}_{wgt} - \mathrm{B})$. Taking variances of both sides, in the limit we have

$V(\hat{U}_N) = \Sigma_U \approx \left[\frac{\partial U_N(B)}{\partial B}\right] V(\hat{b}_{wgt}) \left[\frac{\partial U_N(B)}{\partial B}\right]$. Inverting (provided that $\frac{\partial U_N(\mathrm{B})}{\partial \mathrm{B}}$ is of full rank) gives:

$$V(\hat{b}_{wgt}) \approx \left[\frac{\partial U_N(\mathrm{B})}{\partial \mathrm{B}}\right]^{-1} \Sigma_U \left[\frac{\partial U_N(\mathrm{B})}{\partial \mathrm{B}}\right]^{-1}. \tag{14}$$

The term $\frac{\partial U_N(\mathrm{B})}{\partial \mathrm{B}} = X_N' X_N$ can be estimated using $x'wx$. Note that Σ_U is the variance-covariance

matrix for the k estimated totals of form $\hat{U}^{(k)} = \sum_{c=1}^{a}\sum_{i=1}^{b_c} w_{ci} x_{ci}^{(k)} e_{ci}$ for covariate k, which is estimated

applying (11) and using the regression residuals as estimates of ε. That is:

[28] Our exposition is based on Binder's (1983) approach for implicitly defined estimators. His formulation applies to estimators with a closed form solution (such as regression coefficients) as a special case, but it also applies to other models that economists commonly estimate using maximum likelihood techniques such as logit, probit and Cox proportional hazard models.

$$\hat{\Sigma}_U = \frac{a}{a-1}\sum_{c=1}^{a}(\hat{u}_c - \bar{\hat{u}})(\hat{u}_c - \bar{\hat{u}})' = \frac{a}{a-1}\sum_{c=1}^{a}\hat{u}_c\hat{u}_c', \text{ where } \hat{u}_c = \sum_{i=1}^{b_c} w_{ci}x_{ci}'e_{ci}.^{29} \qquad (15)$$

Putting these pieces together, the linearization variance estimator can also be expressed as $\hat{V}(\hat{b}_{wgt}) = (x'wx)^{-1}x'w\Lambda xw(x'wx)^{-1}$ where Λ is the block diagonal matrix:

$$\Lambda = \begin{bmatrix} \Lambda_1 & 0 & . & 0 \\ 0 & \Lambda_2 & . & 0 \\ . & . & . & . \\ 0 & 0 & . & \Lambda_a \end{bmatrix}, \quad \Lambda_c = e_c e_c' = \begin{bmatrix} e_{c1}^2 & e_{c1}e_{c2} & . & e_{c1}e_{cb_c} \\ e_{c1}e_{c2} & e_{c2}^2 & . & e_{c2}e_{cb_c} \\ . & . & . & . \\ e_{c1}e_{cb_c} & e_{c2}e_{cb_c} & . & e_{cb_c}^2 \end{bmatrix}$$

and e_{cj} is the residual for the j[th] element from the c[th] cluster. If w=I, this simplifies to $\hat{V}(\hat{b}_{OLS}) = (x'x)^{-1}x'\Lambda x(x'x)^{-1}$.

5.2 Estimators based on a random-effects model

From an economist's perspective the difficulty introduced by clustered sampling for the regression model (2) is that $E(\varepsilon\varepsilon') = \Omega \neq \sigma^2 I$. If clusters are sampled independently, the covariance between observations from different clusters is zero, and the variance-covariance matrix of ε can be restricted to have the form:

$$\Omega = \begin{bmatrix} \Omega_1 & 0 & . & 0 \\ 0 & \Omega_2 & . & 0 \\ . & . & . & . \\ 0 & 0 & . & \Omega_a \end{bmatrix}, \quad \Omega_c = \begin{bmatrix} \sigma_{c1}^2 & \sigma_{c1,c2} & . & \sigma_{c1,cb_c} \\ \sigma_{c1,c2} & \sigma_{c2}^2 & . & \sigma_{c2,cb_c} \\ . & . & . & . \\ \sigma_{c1,cb_c} & \sigma_{c2,cb_c} & . & \sigma_{cb_c}^2 \end{bmatrix} .$$

In this case, the asymptotic variance-covariance matrix of $\hat{b}_{OLS}$ is $V(\hat{b}_{OLS}) = (x'x)^{-1}x'\Omega x(x'x)^{-1}$. Given that the number of parameters in Ω grows faster than the sample size, it is not possible to get a consistent estimator of Ω without assuming some parametric structure. An economist might assume some model for

[29] $\bar{\hat{u}} = 0$ due to the normal equations that define $\hat{b}_{wgt}$: $(x'wx)\hat{b}_{wgt} - x'wy = -x'we = 0$. However, with a stratified sample, the variance would involve a sum of H terms $\sum_{c=1}^{a_h}(\hat{u}_{hc} - \bar{\hat{u}}_h)(\hat{u}_{hc} - \bar{\hat{u}}_h)'$, and the $\bar{\hat{u}}_h$ terms must

the error term that imposes restrictions on the elements of Ω_c, thereby reducing the dimensions of the problem enough to estimate Ω based on that model. The random-effects model of section 3 is one example of such a strategy.

An alternative is to exploit the fact that estimation of $V(\hat{\boldsymbol{b}}_{OLS})$ does not require a consistent estimator for Ω, but rather only for the k-by-k matrix $x'\Omega x$. This is the approach taken by White (1980) in developing a heteroscedasticity-consistent estimator for $V(\hat{\boldsymbol{b}}_{OLS})$. The linearization variance estimator uses $x'\Lambda x$ to estimate $x'\Omega x$, and may be viewed as an extension of the White estimator to the case with stratified, clustered samples.[30] This is a robust estimator in the sense that consistency does not require any particular pattern in the correlation of residuals within a cluster, nor does it require the Ω_c to have the same form for different clusters. This is in contrast to the random-effects model that assumes that error terms are equally correlated within clusters, and that the Ω_c are identical in form across clusters.[31]

An added advantage of the linearization variance estimator is that it does not require modification for use with samples with more than one stage of clustering, whereas the standard random-effects model does. To see this, consider the two stages of selection used for CPS samples. First-stage units are either a county or group of counties; second-stage units are groups of approximately four contiguous housing units. One would not expect the error terms of neighbors to be correlated in the same way as the error terms of two observations from different parts of the same county, but treating PSUs as clusters in the standard random effects model implicitly imposes this assumption. Thus, adapting random-effects models for use with multi-stage clustered data would require going beyond the simple model that statistical packages readily estimate. The linearization estimator does not assume any particular pattern of correlations between observations within a first-stage cluster, and so can be used without modification.

---

be retained because they will not in general equal 0.

[30] This estimator is also known as the Huber-White, sandwich, or robust estimator in other strands of the statistical literature. See, for example Carroll and Ruppert (1988) and Diggle, Liang, and Zeger (1994).

5.3 Replication methods

While the linearization approach has a wide variety of applications, it does require one to derive and program a separate set of partial derivatives for each nonlinear estimator. Replication (or resampling) methods provide an alternative—they require greater computational resources, but less estimator-specific derivation. Comparisons of the performance of the linearization estimator and the two replication methods discussed here (the jackknife and balanced repeated replication) have generally found them to have similar properties.[32]

The general idea of replication methods is to draw repeated subsets of the sample, calculate the estimator for each subset, and then estimate the variance based on how much the estimates vary over the repeated subsamples. The drawing of subsamples differs across replication methods, but for each method subsamples are chosen to preserve the design of the overall sample. For clustered designs, that means that a subsample either includes all sample members from a cluster or it excludes all of them.

The most common techniques are known as the 'jackknife' and 'balanced repeated replication' or BRR. BRR is a more specialized procedure than the jackknife in that it is designed for samples in which two clusters are drawn from each stratum.[33] This is a very common design for clustered samples, and the most common alternative—one cluster per stratum—is also usually treated as if it were a two cluster per stratum design.[34] However, the jackknife is applicable to a wider variety of sample designs. In the two-cluster-per-stratum case, properties of the two methods (and their many variants) have been found to be similar to each other and to the linearization estimator, so the choice between them should depend more on how difficult they are to implement than on statistical considerations.[35]

---

[31] Consistency of the linearization estimator does require that increases in sample size take the form of more clusters rather than of increases in the number of elements selected per cluster.
[32] Rust (1985) provides a survey of both empirical and theoretical comparisons.
[33] BRR has been extended to cases with more than two PSUs per stratum. Grouped BRR involves simply combining PSUs within stratum into two groups, and then carrying out BRR as if each group were a single PSU. There are other more complicated ways of adapting BRR, but these are little used in practice.
[34] This is done by pairing strata and treating the combined strata as if they were one.
[35] See Rao, Wu, and Yue (1992) for a brief discussion. Rao and Wu (1985) compare the properties of these estimators.

For jackknife variance estimators, subsamples are formed by dropping the data from a single cluster in turn, and using the remaining data to form an estimate. The variance of an estimator $\hat{q}$ would then be estimated using:

$$\hat{V}(\hat{q}) = \sum_{h=1}^{H} \frac{a_h}{a_h - 1} \sum_{c=1}^{a_h} (\hat{q}_{(hc)} - \hat{q}^*)^2 ,$$

where $\hat{q}_{(hc)}$ is the estimate based on the subsample that excludes cluster *c* in stratum *h*, and $\hat{q}^*$ is a point estimator of q that uses data from all clusters. In the most common version of the jackknife, $\hat{q}^* = \hat{q}$— the estimator based on the entire sample—but there are other variants.

With BRR, replicate half-samples are selected by dropping the data from *H* clusters, one in each stratum, and calculating $\hat{q}_{(r)}$ based on the remaining half-sample of *H* clusters. One could construct up to 2^H different half-samples with a two-PSU-per-stratum design and H strata, but BRR involves choosing 'balanced' half-samples to reduce the number of replicates needed to approximately H.[36] The variance can be estimated using:

$$\hat{V}(\hat{q}) = \frac{1}{R} \sum_{r=1}^{R} (\hat{q}_{(r)} - \hat{q}^*)^2 ,$$

where R is the number of replicates, though there are several variants on this. As for the jackknife, most commonly $\hat{q}^* = \hat{q}$ based on the full sample, but there are other variants.

Table 5 applies these alternative variance estimators to the human capital earnings functions of our earlier examples. Column (1) reports weighted least squares coefficient estimates and column (2) reports estimates of the associated standard errors base on the assumption of IID errors. The remaining three columns report standard errors produced by the linearization method (column 3) and two replication methods (columns 4 and 5). The results illustrate that the design-based methods used in the last three

[36] The balanced half-samples are defined using an *R*-by-*R* Hadamard matrix—a *k* by *k* matrix whose elements are each either +1 or –1and which satisfies $HH' = kI$, with *k*=1, 2, or a multiple of 4. Replicate *r* is formed by choosing the first PSU in stratum h when H[h,r]=+1, and the second PSU in that stratum when H[h,r]=-1. $\hat{q}_{(r)}$ is

columns yield very similar standard error estimates in these data. The design-based estimates are fairly close to the IID estimates of column (2) for most variables, but the IID estimates are much smaller for the MSA dummy. As before, the MSA dummy is particularly sensitive to clustering because it typically does not vary at all within clusters. Table 5 suggests that it is important to account for the survey design in estimating standard errors, but that the choice between alternative methods is of second order importance. This view is supported by the related statistical literature.

5.4 Practical considerations

Public-use files almost universally include survey weights, so the information required to account for varying probabilities of selection is readily available. Accounting for stratification and clustering requires access to variables that group observations into strata and clusters—information which may or may not be available to the public user. Because this information potentially identifies small geographic areas, it is sometimes suppressed in creating public-use files for household surveys out of concern that, in combination with demographic information, it might compromise respondent confidentiality. As a compromise, many surveys now provide enough information on public-use files to at least approximate design effects.[37] In addition, several federal statistical agencies have instituted programs that allow economists to have restricted access to non-public-use databases that include stratum and cluster identifiers.

Until recently, econometric software packages did not include procedures meant to handle survey data, but this has also begun to change. Procedures for regression and relatively simple non-linear

---

formed by doubling all of the weights, and then applying estimator $\hat{\boldsymbol{q}}$ to the half-sample. R is the smallest multiple of four such that the number of strata is less than or equal to R ($H \le R \le H + 3$).

[37] For example, the PSID, the National Health Interview Survey, and the Health and Retirement Survey each provide users with variables that can be used to group observations into something approximating strata and PSUs for the purpose of computing standard errors. The restricted-release Geocode version of the NLSY79 also provides such codes. The CPS does not provide such codes on public-use files.

estimators (e.g. logit and proportional hazard models) are now included in several widely used packages.[38]

6. Conclusions

Survey data are a staple of econometric analysis, but economists are often only vaguely familiar with how survey samples are designed. This is unfortunate, as in some cases the failure to account for a survey's design and implementation can result in inference procedures that are quite inaccurate. To summarize, failure to account for unequal probabilities of selection can lead standard point estimators to be biased, and failure to account for clustering can lead to severely understated variance estimates. Put together, the failure to account for survey design can lead to inaccurate inferences being drawn from survey data.

There is an increased sensitivity to the effects of survey design on inference within the econometrics community, as is evident from the large number of recent empirical papers that report accounting for survey design effects, and from the small but growing econometric literature on the effects of survey design (e.g., Wooldridge, 1998 and 1999; Imbens and Lancaster, 1996; Hellerstein and Imbens, 1999). We hope that this paper serves as a useful introduction to the issues involved, and as a practical introduction to the choices that applied economists need to make when they use survey data.

[38] www.fas.harvard.edu/~stats/survey-soft/survey-soft.html maintains a list of software packages that include design-based estimators.

References

Binder, David. 1983. "On the Variances of Asymptotically Normal Estimators From Complex Surveys," International Statistical Review, vol. 51, pp. 279-292.

Binder, David. 198?. "Use of Estimating Functions for Estimation from Complex Surveys." Journal of the American Statistical Association, vol. 89, pp. 1035-43.

Carroll, R.J. and Ruppert, D. 1988. Transformation and Weighting in Regression. London: Chapman and Hall.

Cochran, William G. 1977. Sampling Techniques. 3rd edition. New York: John Wiley & Sons.

Deaton, Angus. 1997. The Analysis of Household Surveys: A Microeconometric Approach to Development Policy. Baltimore: The Johns Hopkins University Press.

Diggle, P.J., Liang, K.-Y., and Zeger, S.L. 1994. Analysis of Longitudinal Data. Oxford: Clarendon Press.

DuMouchel, William H., and Duncan, Greg J. 1983. "Using Sample Survey Weights in Multiple Regression Analyses of Stratified Samples." Journal of The American Statistical Association, vol. 78, no. 383, pp. 535-543.

Frankel, Martin R., Harold A. McWilliams, and Bruce D. Spencer. 1983. National Longitudinal Survey of Labor Force Behavior, Youth Survey: Technical Sampling Report. Chicago: National Opinion Research Center.

Freeman, Richard B. 1986. "Demand for Education." In "Handbook of Labor Economics, volume 1, edited by Orley Ashenfelter and Richard Layard. Amsterdam: North-Holland.

Fuller, Wayne A. 1984. "Least Squares and Related Analyses for Complex Survey Designs." Survey Methodology, vol. 10, no.1, pp. 97-118.

Greene, William . 1997. Econometric Analysis, 3rd Edition. New York: Prentice Hall.

Hausman, Jerry A. 1978. "Specification Tests in Econometrics." Econometrica, vol. 46, pp. 1251-1271.

Hausman, Jerry A., and David A. Wise. 1981. "Stratification on Endogenous Variables and Estimation: The Gary Income Maintenance Experiment." In Structural Analysis of Discrete Data with Econometric Applications, ed. By Manski, C.F. and McFadden, D. Cambridge, Massachusetts: MIT Press.

Hellerstein, Judith, and Guido Imbens. 1999. "Imposing Moment Restrictions from Auxiliary Data by Weighting." Review of Economics and Statistics, vol. 81, no. 1, pp. 1-14.

Huber, P.J. 1967. "The Behavior of Maximum Likelihood Estimates Under Non-Standard Conditions." In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. Berkeley, CA: University of California Press, vol. 1, pp. 221-233.

Imbens, Guido W., and T. Lancaster. 1996. "Efficient Estimation and Stratified Sampling." Journal of Econometrics, vol. 74, pp. 289-318.

Kott, Phillip S. 1991. "A Model-Based Look at Linear Regression With Survey Data." The American Statistician, vol. 45, no. 2, pp. 107-112.

Lillard, Lee, James P. Smith, and Finis Welch. 1986. "What Do We Really Know about Wages? The Importance of Nonreporting and Census Imputation." Journal of Political Economy, vol. 94, no. 3, part 1, pp. 489-506.

Little, R.J.A. and D.B. Rubin. 1987. Statistical Analysis with Missing Data. New York: Wiley.

MaCurdy, Thomas, Thomas Mroz, and R. Mark Gritz. 1998. "An Evaluation of the National Longitudinal Survey of Youth." Journal of Human Resources, vol. 33, pp. 345-436.

Moulton, Brent R. 1986. "Random Group Effects and the Precision of Regression Estimates." Journal of Econometrics, vol. 32, pp. 385-397.

Moulton, Brent R. 1990. "An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Units." Review of Economics and Statistics, vol. 72, pp. 334-38.

Pfeffermann, D. 1993. "The Role of Sampling Weights when Modeling Survey Data." International Statistical Review, v. 61, no. 2, pp. 317-337

Pfeffermann, D. 1996. "The Use of Sampling Weights for Survey Data Analysis." Statistical Methods in Medical Research, vol. 5, pp. 239-61.

Rao, J.N.K., and Wu, C.F.J. 1985. "Inference from Stratified Samples: Second-Order Analysis of Three Methods." Journal of the American Statistical Association, vol. 80, pp. 620-630.

Rao, J.N.K., Wu, C.F.J., and Yue, K. 1992. "Some Recent Work on Resampling Methods for Complex Surveys." Survey Methodology, vol. 18, pp. 209-217.

Rust, Keith. 1985. "Variance Estimation for Complex Estimators in Sample Surveys." Journal of Official Statistics, vol.1, pp. 381-397.

Scott, A.J. and Holt, D. 1982. "The Effect of Two-Stage Sampling on Ordinary Least Squares Methods." JASA, vol. 77, pp. 848-854.

Shao, J. 1996. "Resampling Methods in Sample Surveys" (with discussion). Statistics, vol. 27, pp. 207-54.

White, Halbert. 1980. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." Econometrica, vol. 48, pp. 817-38.

Wolter, Kirk M. 1985. Introduction to Variance Estimation. New York: Springer-Verlag.

Wooldridge, Jeffrey M. 1998. "Asymptotic Properties of Weighted M-Estimators for Standard Stratified Samples." Working Paper, Michigan State University, April 1998.

Wooldridge, Jeffrey M. 1999 "Asymptotic Properties of Weighted M-Estimators For Variable Probability Samples." Econometrica, vol. 67, pp. 1385-1406.

Table 1: OLS and Design-Based Estimates of Log Wage Regression

Independent Variable	(1) OLS/IID Estimates	(2) Design-based Estimates
Black	-.063	-.102
	(.014)	(.017)
Hispanic	-.048	-.080
	(.012)	(.015)
Female	-.313	-.303
	(.008)	(.009)
Age	.0092	.0097
	(.0004)	(.0005)
Years of schooling	.0972	.0984
	(.0015)	(.0018)
Intercept	.866	.845
	(.028)	(.035)

Notes: The dependent variable is the log of average hourly earnings. Data are from the March 1997 CPS. Standard errors are in parentheses.

Table 2: Design Variables for Selected Large U.S. Surveys

(1) Stratification	(2) Probabilities of selection	(3) Nonresponse adjustments	(4) Poststratification	(5) Clustering
A. Surveys of households				
1. *Current Population Survey*: Measures national and state unemployment rates.				
Geographic areas (PSUs) stratified by state, number of unemployed by gender, number of female-headed families, fraction of housing units with >=3 persons, employment and wages by industry from BLS	State, Hispanic status of dwelling unit in earlier rotation group (March only)	State, MSA status and size, central city/not central city or rural/ nonrural	Black/non-black census population within state (for non-self-representing PSUs only); national age/sex /race and age/sex/ Hispanic population; state population	(1) County or county group; (2) groups of about 4 dwelling units. One cluster selected per noncertainty stratum.
2. *National Longitudinal Survey of Youth, 1979*: Measures youth labor market dynamics by race, income status, sex.				
Geographic areas (PSUs) stratified by region, SMSA status, county size, percent black, median family income	Race of youth and 1977-78 family income as reported in screener; characteristics of 3rd stage unit of selection (see last column): percent black, percent Hispanic, percent low income	Completion rate in 3rd stage sampling unit for screening interview; whether complete income information given in screener; race/ethnicity, sex, and birth year within PSU	Population by race/ethnicity, sex, and birth year based on Census Bureau estimates.	(1) County or county group; (2) block group or enumeration district; (3) small area with 100+ dwelling units. One cluster selected per noncertainty stratum.
3. *Health and Retirement Survey*: Measures dynamics of retirement and aging.				
Geographic areas (PSUs) stratified by MSA status, population of area, geographic location	Racial/ethnic composition of area, marital status, number of age-eligible persons in household, age, Florida	PSU, racial/ethnic composition of neighborhood (within PSU)	Number of households by region by race by marital status; population by region by race/ethnicity by sex by age group	(1) MSA, county or county group; (2) blocks or block groups. One cluster selected per noncertainty stratum.

Table 2: Design Variables for Selected Large U.S. Surveys (Continued)

(1) Stratification	(2) Probabilities of Selection	(3) Nonresponse adjustments	(4) Poststratification	(5) Clustering
B. Surveys of establishments				
1. *Annual Survey of Manufacturers*: Measures shipments, etc. by industry, and for industry by state.				
Industry	Value and time-series variability of product class shipments	No adjustment made to weights—use imputation to adjust for nonresponse	No adjustment of weights, but published estimates include an adjustment equal to difference (in census year) between ASM and Census of Manufactures estimates	No clustering of establishments, but if product line is unit of analysis, cluster= establishment
2. *Medical Expenditure Panel Survey IC*: Measures % of establishments offering health insurance, % of workers covered, by state.				
State; firm and establishment size	State; firm and establishment size; number of establishments in firm	Whether establishment was known to offer insurance; industry; single vs. multiunit; state; firm and establishment size	Employment by (state by firm size by establishment size) from Census Bureau's business list	No clustering of establishments, but if health plan is unit of analysis, cluster= establishment
3. *Employer Cost Index*: Provides basis for index of changes in employee compensation for a fixed bundle of jobs.				
Industry	Establishment and occupational employment	Industry and establishment size	Weights not adjusted but published estimates of compensation levels poststratified to employment counts by industry from Current Employment Statistics survey	No clustering of establishments, but if occupation is unit of analysis, cluster= establishment

Notes: The NLSY79 includes an equal-probability-of-selection component of the sample, along with other subsamples that oversample certain populations. The description of variables determining probabilities of selection for the NLSY79 pertain to the sample as a whole, or to the special subsamples. Descriptions of variables used for nonresponse adjustments do not include all adjustments for attrition in panel surveys.

Table 3: Effects of Weighting When Controlling for Design Variables

	Main Design Variables NOT Included on Right-Hand Side		Main Design Variables ARE Included on Right-Hand Side	
	(1) Unweighted	(2) Weighted	(3) Unweighted	(4) Weighted
1. Black	-.063	-.102	-.111	-.120
	(.014)	(.017)	(.014)	(.017)
2. Hispanic	-.049	-.080	-.090	-.112
	(.012)	(.015)	(.013)	(.016)
3. Female	-.313	-.303	-.311	-.301
	(.008)	(.009)	(.008)	(.009)
4. Age	.0092	.0097	.0091	.0096
	(.0005)	(.0005)	(.0005)	(.0005)
5. Years of schooling	.097	.098	.094	.096
	(.0017)	(.0018)	(.0016)	(.0018)
6. State dummies?	No	No	Yes	Yes

Notes: Data come from the March 1997 CPS. The first two columns of estimates are repeated from Table 1, except that standard error estimates account for stratification and clustering in all columns.

Table 4: Examples of Misspecification Effects From a Log Wage Regression

Independent variables	(1) $\hat{r}_x$	(2) Misspecification factor based on random components model	(3) Misspecification factor based on linearization estimator
CPS, March 1997: $\hat{r}_e$ =.020			
Female	-.010	1.00	1.07
Age	.018	1.01	1.14
Years of schooling	.061	1.04	1.10
Black	.128	1.18	0.94
Hispanic	.247	1.49	1.04
MSA	.981	1.49	1.24
NLSY79: $\hat{r}_e$ =.046			
Female	.001	1.00	0.85
Age	.011	1.01	1.01
Years of schooling	.074	1.05	1.05
Black	.548	1.47	1.13
Hispanic	.541	1.44	1.56
MSA	.420	1.32	1.28

Notes: Column 2 gives the square root of the diagonal elements of $I + (\Lambda - I)\rho_e$, as defined in text. In the CPS, sample PSUs have on average 47.3 persons in the regression sample. In the NLSY79, sample PSUs have on average 45.5 persons in the regression sample. Column 3 gives the ratio of linearization standard errors to standard errors based on the estimator $(X'X)^{-1}\hat{s}^2$. Noncertainty strata were paired up to approximate a 2-PSU-per-stratum design.

Table 5: Alternative Design-Based Standard Error Estimates

		Alternative Estimates of Standard Errors			
	(1)	(2)	(3)	(4)	(5)
Independent variables	Coefficients	IID	Linearization	BRR	Jackknife
CPS, March 1997					
Female	-.301	.00774	.00863	.00865	.00863
Age	.0098	.00042	.00054	.00054	.00054
Years of schooling	.0933	.00150	.00159	.00160	.00159
Black	-.126	.01243	.01648	.01654	.01649
Hispanic	-.122	.01388	.01433	.01423	.01433
MSA	.256	.01012	.01319	.01328	.01317
NLSY79, 1994 Wave					
Female	-.288	.01353	.01656	.01671	.01665
Age	.017	.00290	.00370	.00375	.00371
Years of schooling	.088	.00262	.00393	.00395	.00394
Black	-.170	.02259	.02339	.02355	.02354
Hispanic	-.021	.03286	.02754	.02805	.02730
MSA	.231	.01648	.02986	.03018	.03013

Notes: For the CPS, the 531 strata used in earlier tables are grouped into 100 grouped-strata to accommodate constraints in the software we used to estimate the BRR and jackknife standard errors.